LEXiTRON Dictionary Improvement based on the Computational Lexicography Method

Sawittree Jumpathong[†a)], Sitthaa Phaholphinyo[†], Dhanon Leenoi[†],
Kanyanut Kriengket[†], Prachya Boonkwan[†], and Thepchai Supnithi[†]

Language and Semantic Technology Laboratory,
National Electronics and Computer Technology Center, Thailand

***Abstract*— Traditional methods of dictionary construction are time-consuming and labor-intensive, due to the immense burden of manual data compilation and management. In this paper, we propose to use the Computational Lexicography method to abridge the dictionary development. We develop a semi-automatic dictionary-making procedure that applies the Computational Lexicography method. Apart from extending LEXiTRON to 100,000 entries, we also incorporate multiple word expressions, example sentences, and additional phonetic transcription. Especially in phonetic transcription, we propose a semi-automatic method to reduce the burden of transcribing all new words manually. With the benefit of the Computational Lexicography method, a bunch of English entries which should be done by the cognoscenti are accomplished in the limited time, 2.3 times faster than the classical method.**

***Index Terms*— Lexicography, Electronic Dictionary, Natural Language Processing**

I. Introduction

Dictionary is a crucial component of natural language processing. It provides lexical information for various NLP tasks such as part-of-speech tagging, syntactic parsing, machine translation, and information retrieval. Dictionaries are, on the other hand, hard to construct because the traditional development process is laboriously manual and time-consuming, resulting in the lack of linguistic resources for less-privileged languages.

For English-Thai translation and vice versa, LEXiTRON is one of the most popular electronic dictionaries with a two decades long history of development. LEXiTRON is revised and updated from time to time by comparing it with three best-seller-dictionaries in the market: (1) SE-ED's English-Thai Dictionary with Idioms & Phrases, (2) Times-Chambers Learner's Dictionary (English-Thai), and (3) New Model English-Thai Dictionary. Some thorough comparison helps us indicate the lack in our dictionary. To beat those three dictionaries, the lexical entries must be increased to at least 100,000 entries together with the Thai transcriptions, multi-word expressions and example sentences.

II. Dictionary Construction

This section will elaborate the notions of semantics and lexicography, as they lay a rigid foundation for dictionary construction.

a. Semantics

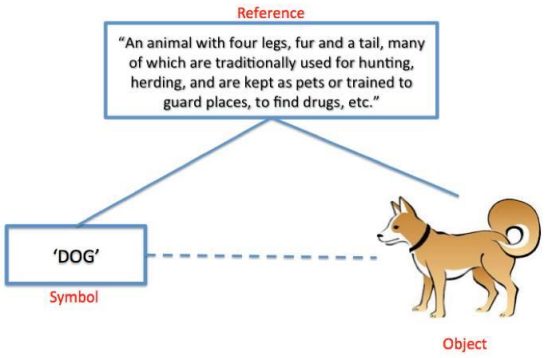

Fig. 1. The relation between symbol, reference, and object.

Semantics is the study of meaning [9, 10, 14] conveyed through language. A word, i.e. symbol, is indirectly related to the object of which presenting via conceptual reference [11].

In Figure 1, symbol 'dog' associates directly to conceptual reference 'an animal with four legs, fur and a tail, many of which are traditionally used for hunting, herding, and are kept as pets or trained to guard places, to find drugs, etc.', then linking to the object. Conversely, the word 'dog' indirectly connects with 'the object', perceptible by one or more of the senses especially by vision or touch [12].

The reference to be composed as a definition can be classified into two main elements [1, 7]: genus (common attributes) and differentia (specific attributes). Moreover, encyclopedic knowledge (additional information acquired from world knowledge) [10] can optionally be included (see Table I).

[†]The authors are with National Electronics and Computer Technology Center, 112 Phahon-Yothin Rd., Klong-Luang, Pathumthani, 12120, Thailand.
a) sawittree.jumpathong@nectec.or.th

TABLE I
Genus, Differentia, and Encyclopedic Knowledge.

semantic element	'POLICEMAN'	'SOLDIER'
Genus	Government employee	Government employee
Differentia	Trained in method of law enforcement, crime prevention and detection.	Trained in method of battle to prevent a population or area from being captured or occupied by enemies.
Encyclopedic knowledge	[French, from Old French policie, civil organization, from Late Latin polītīa, from Latin, the State, from Greek polīteia, from polītēs, citizen, from polis, city.]	[Middle English soudier, mercenary, from Anglo-Norman soudeour, soldeier and Old French soudoior, soudier, both from Old French sol, soud, sou, from Late Latin solidum, soldum, pay.]

Applied from the Free Dictionary

b. *From Conceptual Reference to Definition*

A definition is manually composed with respect to the aforementioned elements. Genus and differentia together with encyclopedic knowledge are juxtaposed respectively in brief and free from confusion or doubt. For example, 'police – a government employee who is trained in method of law enforcement, crime prevention and detection; from Old French policies, 'civil organization'.

c. *Lexicography*

Lexicography is the crafting of dictionary. There are two approaches of lexicography: expertise-based approach and corpus-based approach [3, 4]. The first is to compile on the basis of the cognoscenti (domain experts), while the latter is to build from the corpus which records language usage and reflects the linguistic phenomena.

Moreover, the methods of the dictionary making can be distinguished to two techniques: manual and semi-automatic. Comparing with other dictionaries in Thailand's market, only LEXiTRON has been compiled through the corpus-based approach and the semi-automatic technique.

III. System Overview

a. *LEXiTRON Dictionary Extension*

LEX*i*TRON was constructed using the corpus-based approach and the semi-automatic technique. Pursuing the practical language use, we take into account two linguistic sources: a large text corpus based on general-domain news and emerging vocabulary frequently searched by users. This results in a tremendous burden of manual lexicography in which lexicon selection, translation definition, and dictionary compilation have to take place. To facilitate this process, we propose a system design that incorporates both sources and automates the workflow of dictionary construction, as illustrated in Fig. 2.

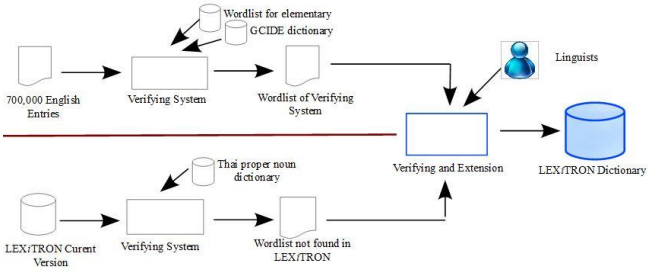

Fig. 2. LEX*i*TRON Dictionary Extension process

There are two sources of vocabulary. First, we chose words from 700,000 English words queried from LEXiTRON by its members but not found. The words were verified with two references: (1) elementary English word list for primary school readers [16] and (2) GCIDE dictionary [5] consisting of 131,565 headwords. Second, we chose words from Thai proper noun dictionary created by our linguists, and compared it against the current version of LEXiTRON (containing 83,843 entries). Afterwards, the words not found in LEXiTRON were extracted from our dictionary of Thai proper nouns.

As a result, we selected 16,157 entries from the aforementioned process. The English words for the secondary readers and the high-frequently searched words were selected to compose Thai definitions, while the academic and rarely used words are postponed. Then, we composed their Thai definition and example sentences, and enrich them with linguistic features, such as plural form, irregular verb form, and comparison form.

b. *Semi-automatic produce-and-predict Thai transcriptions system*

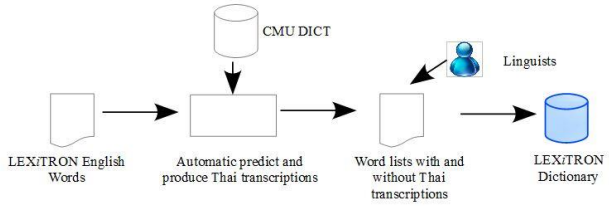

Fig. 3. The procedure of produce and predict Thai transcriptions

One challenging step in corpus-based dictionary construction is to provide phonetic transcription for a voluminous amount of vocabulary. As seen in Fig 3, we automatically look up each word in CMUDICT and retrieve its IPA pronunciation. We then transcribe it into Thai using the mapping table in Table II, and we have the result validated by the linguists.

In the case where pronunciations are not found in CMUDICT [2], our reference dictionary, we predict and produce the most likely pronunciation. We first treat the input word as a compound word and attempt to decompose

it into subunits existing in the reference dictionary. For instance, the word 'aftertime' can be segmented into subunits 'after' and 'time'. We enumerate all combinations of subunits that are found in reference dictionary and produce a list of pronunciation candidates in Thai. If subunits are not found, we compute the orthographically closest word in the reference dictionary using the Dynamic Programming paradigm. For example, “arbored” is orthographically close to “harbored”. Finally, the linguists validate and handpick the correct ones from the provided list. Table III shows the final results of pronunciation generation.

TABLE II
Mapping between English and Thai Phonemes

Type	CMU Pronunciations	IPA	Thai Transcriptions
1. Consonant	B	b	บ
	D	d	ด
	F	f	ฟ
	M	m	ม
	N	n	น
	R	r	ร
	SH	ʃ	ช
	T	t	ท
2. Vowel	AA	ɑ	ออ, อา
	AE	æ	แอ
	AH0	ə	เออะร์
	AH1	ʌ	อะ
	AY	aɪ	ไอ
	ER	ɚ	เออะร์/เออร์
	EY	eɪ	เอ
	IH	ɪ	อิ
3. Stress	0	-	-
	1	ˈ	ˈ
	2	ˌ	ˌ

TABLE III
Result of Thai Transcription

Vocabulary	CMU Pronunciations	IPA	Thai Transcriptions
information	IH2 N F ER0 M EY1 SH AH0 N	ˌɪnfɚˈmeɪʃən	ˌอินเฟอะร์ˈเมเชิน
aftertime	AE1 F T ER0 T AY1 M	ˈæftɚˈtaɪm	ˈแอฟเทอะร์ˈไทม
arbored	AA1 R B ER0 D	ɑrbɚd	อาร์เบอร์ด

IV. Results

In this section, we describe the results of dictionary construction using our method. It is, however, cumbersome to compare our method against the traditional method, because the construction of market dictionaries is proprietary and confidential. We instead quantify our workload in teams of man-day measure and accuracy.

a. Choosing words

According to the verify words process (section 3.2), when 700,000 entry not found in LEX*i*TRON were verified with two reference. We found that the 700,000 entries found in wordlist for elementary are 435 entries. The 700,000 entries found in GCIDE dictionary are 21,800 entries.

Furthermore, we compared Thai proper noun dictionary with LEX*i*TRON current version. We found that words of Thai proper noun dictionary not found in LEX*i*TRON are 13,000 entries.

b. Accuracy of the transcription system

The accurate output level from the *produce* method is 93 percent. The *predict* method makes 75, 60 percent accuracy in compound and isolated word respectively. The detail was shown in Table IV.

TABLE IV
Thai Transcription Construction Method

Thai transcriptions construction method	Amount (words)	Accuracy	
		words	%
1. Produce method (automatic)	73,235	68,108	93.00%
2. Predict method (automatic) 2.1 Compound word 2.2 Isolated word	 16,351 2,235	 12,263 1,341	 75.00% 60.00%
3. Transcribe manually by Linguists (vocabulary beyond 1 and 2)	8,179	-	-
Total	100,000	-	-

c. Timeline comparison between corpus-based approach and expertise approach

We compare time spent in dictionary construction procedure between two approaches in Table V. The results show that developing dictionary with corpus-based approach is faster than the expertise approach for 2.3 times. Particularly, vocabulary selection and Thai transcriptions construction step in this approach is faster than expertise approach for 1.7 times and 3.7 times, respectively.

TABLE V
Comparison between Two Corpus-based and Expertise Approach

Dictionary construction procedure	Amount of Worker (person)	Period of time (day)	
		manual	semi-automatic
Vocabulary selection (16,157 entries)	6	71	42
Thai transcriptions 100,000 entries	7	139	38
Grammatical derivation extraction	2	3	1
Spelling styles creation (American English and British English)	1	2	2
Illustration insertion	1	15	15
Head word and sub head word assigning	1	5	2
Vocabulary gathering	1	2	2
Total	**18**	**237**	**102**

d. Overall User Satisfaction

We assess the overall user satistisfaction scores with the LEX*i*TRON electronic dictionary users, most of which being in the linguistic field. They were experts and lecturers from any universities, e.g. Chulalongkorn University, Silpakorn University, and Ramkhamhaeng University, in order to exam language correctness and usage. They provided the product assessments of overall dictionary and dictionary data from 1 to 5 scores which 5 was the most satisfaction, and their results were as the Table VI.

TABLE VI
Assessment of LEXiTRON Dictionary

Criteria		Average Satisfaction Scores
1. LEX*i*TRON dictionary		
1.1	Numbers of entries	4.6
1.2	Data completeness	4.2
1.3	Data correctness	4.0
1.4	Dictionary Usage	4.4
1.5	Suitability of student dictionary	4.8
1.6	Specific dictionaries support	4.4
Total average		**4.4**
2. LEX*i*TRON data		
2.1	Translation	4.8
2.2	Part of speech	4.8
2.3	Pronunciation by Thai alphabets	4.6
2.4	Sample sentences	3.6
2.5	Additional data, i.e. domain, verb forms, pictures, idioms	4.0
Total average		**4.4**

According to the above table, users were very satisfied with LEX*i*TRON dictionary in both general qualities and dictionary data. For general qualities, they were pleased with all criteria: numbers of entries, data completeness, data correctness, dictionary usage, suitability of student dictionary, and specific dictionaries support. As can be seen, the total averages of satisfaction scores were nearly proximate. Besides, regarding the dictionary data, they were quite satisfied with translation, part of speech, and pronunciation by Thai alphabets, sample sentences, and any additional data. As can be seen, the total averages of satisfaction scores were quite high. LEX*i*TRON dictionary was quantitatively and qualitatively satisfactory especially in its dictionary data. Hence, LEX*i*TRON receives high overall user satisfaction (see Table VI), suggesting that it is suitable for daily usage.

V. Discussion

a. Semi-automatic Thai Transcription System

In section 4.2, the proposed method significantly reduces the burden of manual transcription in dictionary construction. It can be seen that we obtain 91.82% coverage of the new vocabulary by using CMUDICT and the predict method. The total accuracy of the automatic transcription on this coverage is 88.99%. Only 18,287 entries are corrected and manually transcribed, as opposed to doing so on the entire 100,000 entries in the traditional method. This shows that our semi-automatic method increases the productivity of dictionary construction.

b. Period of time

The result shows that developing a dictionary with corpus-based approach, LEX*i*TRON in our case, is 2.3 times faster than expertise approach. Particularly, vocabulary selection and Thai transcriptions construction steps in this approach is faster than expertise approach for 1.7 times and 3.7 times, respectively.

One topic to address here is our semi-automatic method significantly shortens the time duration for manual transcription by 3.7 times — from 139 days to 38 days. This process is expensive in the traditional method because it requires a labor force to tackle a large amount of vocabulary, resulting in exorbitant expenditure. Therefore, our method makes dictionary construction much more economical and less labor-intensive.

c. Overall User Satisfaction

According to the assessments of overall user satisfaction, users in the linguistic field from any universities were satisfied with LEX*i*TRON electronic dictionary. On average, the satisfaction scores of overall LEX*i*TRON dictionary and data was high at 4.4 and 4.4, respectively. The experts suggested us improve the current dictionary with more sample sentences, comparison of confusing word usage, e.g. make/do, much/many, little/few, a lot/plenty, and on time/in time.

It is also worth mentioning what LEXiTRON still lacks. In Table VI, data completeness and correctness obtain the least satisfaction scores (4.2 and 4.0, respectively). When we further analyze the feedback, we found that most users require additional sample sentences and annotated information e.g. domain, verb forms, pictures, and idioms. We understand that our users assessed our system based on their academic perspective, in which students make use of the resultant dictionary.

d. Observing the Remainders

In the verification process, all 700,000 entries of LEXiTRON's search log were verified with GCIDE dictionary. We found that 21,800 entries of them exist in GCIDE dictionary. We analyze the remaining entries not found in GCIDE dictionary. We found that the majority of them are misspelled words (66.93%) and the remainders are phrases and sentences (33.07%).

VI. Conclusion and Future Work

We have presented the semi-automatic approach of dictionary construction based on computational lexicography method where automatic word transcription is included to further reduce the time and labor capitals. It is clear that the

dictionary making process through this method is faster 2.3 times vis-à-vis the classical approach.

Our future work remains as follows. 700,000 unknown English entries from users, not found in the dictionary, should be further analyzed. The findings reflecting the linguistic phenomena could improve the database so as to extend the dictionary entries as well as additional information afterwards.

References

1. Atkins, B.T. Sue and Rundell, Micheal: The Oxford Guide to Practical Lexicography. Oxford University Press, Oxford (2008)
2. CMU Pronouncing Dictionary (November 2015). http://www.speech.cs.cmu.edu/cgi-bin/cmudict
3. Dhanon Leenoi: The Construction of Thai WordNet of 1st Order Entity Common Base Concepts using a Bi-directional Translation Method and with Dictionaries of Different Compilational Approaches. M.A. Thesis, Chulalongkorn University, Bangkok (2009)
4. Dhanon Leenoi, Thepchai Supnithi, and Wirote Aroonmanakun: Building a Gold Standard for Thai WordNet. In: Proceeding of the International Conference on Asian Language Processing 2008, pp 78 – 82. COLIPS, Singapore (2008)
5. GCIDE dictionary (November 2015). http://www.ibiblio.org/webster/
6. Kanittha Navarat, et. al.: Times-Chambers Learner's Dictionary (English-Thai). Samakkhisan (Dokya), Bangkok (1997)
7. Landau, Sidney I.: Dictionaries: The Art and Craft of Lexicography. Cambridge University Press, Cambridge (1989)
8. Lyons, J.: Semantics. Cambridge University Press, Cambridge (1977)
9. Leech, Geoggrey N.: Semantics. Penguin, Harmondsworth (1983)
10. Nida, Eugine Albert: Componential Analysis of Meaning: An Introduction to Semantic Structures (Approaches to Semiotics). Mouton, Paris (1975)
11. Ogden, C.K. and Richards, I.A.: Meaning of Meaning. Mariner Books, New York (1989)

Sawittree Jumpathong received B.Sc. in Computer Science from Naresuan University in 2008. She has been with Language and Semantic Technology Lab at NECTEC in Thailand. Her topics of interest include: semantics, Software engineering, and natural language processing.

Sitthaa Phaholphinyo received B.A. in Linguistics from Thammasat University in 1997. He received M.A. in Language and Communication from National Institute of Development Administration in 2007. Since 2000, he has been with Language and Semantic Technology Lab at NECTEC in Thailand. His topics of interest include: sociolinguistics, semantics, pragmatics and syntax.

Dhanon Leenoi received B.A. from Thammasat University and M.A. from Chulalongkorn University. Since 2009, he has been with Language and Semantic Technology Laboratory at NECTEC, Thailand. His topics of interest include: computational linguistics, cognitive linguistics, lexicography, semantics and pragmatics.

Kanyanut Kriengket received B.A. (Second Class Hons.) and M.A. degrees in Thai Language from Chulalongkorn University in 2003 and 2008, respectively. Since 2008, she has been with Language and Semantic Technology Lab at NECTEC in Thailand. Her topics of interest include: semantics, morphology, syntax, language learning, corpus linguistics, and natural language processing.

Prachya Boonkwan received B.Eng. and M.Eng. degrees in Computer Engineering from Kasetsart University in 2002 and 2005, respectively. He received a Ph.D. degree in Informatics from the University of Edinburgh, UK, in 2014. Since 2005, he has been with Language and Semantic Technology Lab at NECTEC in Thailand. His topics of interest include: grammar induction, statistical parsing, statistical machine translation, natural language processing, machine learning, and formal syntax.

Thepchai Supnithi received a B.S. degree in Mathematics from Chulalongkorn University in 1992. He received M.S. and Ph.D. degrees in Engineering from Osaka University, Japan, in 1997 and 2001, respectively. Since 2001, he has been with Language and Semantic Technology Lab at NECTEC in Thailand. His topics of interest include: statistical machine translation, natural language processing, machine learning, ontology, and knowledge engineering.